Preference Ranking Optimization for Human Alignment

Abstract

Large language models (LLMs) often contain misleading content, emphasizing the need to align them with human values to ensure secure AI systems. Reinforcement learning from human feedback (RLHF) has been employed to achieve this alignment. However, it encompasses two main drawbacks: (1) RLHF exhibits complexity, instability, and sensitivity to hyperparameters in contrast to SFT. (2) Despite massive trial-and-error, multiple sampling is reduced to pair-wise contrast, thus lacking contrasts from a macro perspective. In this paper, we propose Preference Ranking Optimization (PRO) as an efficient SFT algorithm to directly fine-tune LLMs for human alignment. PRO extends the pair-wise contrast to accommodate preference rankings of any length. By iteratively contrasting candidates, PRO instructs the LLM to prioritize the best response while progressively ranking the rest responses. In this manner, PRO effectively transforms human alignment into aligning the probability ranking of n responses generated by LLM with the preference ranking of humans towards these responses. Experiments have shown that PRO outperforms baseline algorithms, achieving comparable results to ChatGPT and human responses through automatic-based, reward-based, GPT-4, and human evaluations.